Foundational Aspects of Schema Mapping Optimization and Normalization

Author: Vadim Savenkov
Reviewer and Supervisor: Prof. Dr. Reinhard Pichler
External reviewer: Prof. Dr. Nicole Schweikardt

Abstract

Schema mappings represent a basic concept of data integration and data exchange, expressing the relationship between database schemas. They can be seen as declarative programs, transforming the data from one schema to another, or rewriting a query against one schema as a query against another schema. A typical situation is when syntactically different mappings have the same effect with respect to a given task in information integration. However, syntactical differences may have dramatic effects on the consumption of computational resources, which are required for solving the task.

Considering mappings leading to the same end result for the task at hand as equivalent, one can formulate the semantic optimization problem as follows: among multiple equivalent mappings find those that yield the least resource consumption. The first step towards the goal of schema mapping optimization has been made by Fagin et al. in 2008 by introducing three basic notions of schema mapping equivalence and studying their properties. This dissertation extends the theory of schema mapping optimization in several directions.

The first part of this dissertation deals with the question of optimization and normalization of schema mappings with respect to the standard notion of logical equivalence. We formulate the concrete optimality criteria for schema mappings given by embedded dependencies, and define a system of rewrite rules, transforming a mapping given by a set of source-to-target dependencies into an optimal one. We also prove that the result of applying our rewrite rule system is unique up to renaming of variables. Moreover, we extend this result by defining a unique normal form in the presence of target equality generating dependencies and show the trade-off between uniqueness and optimality in this setting.

In the second part of the dissertation, we move on to the relaxed notions of equivalence, proposed by Fagin et al.: namely, to data exchange equivalence and to conjunctive query equivalence. If no integrity constraints are defined over the target schema, these notions are known to coincide with logical equivalence. We show that this result holds even if the target schema includes integrity constraints, under the assumption that the schema and the constraints are fixed. If the target constraints are taken as part of the mapping and are allowed to vary, the relaxed notions of equivalence are known to become undecidable. For conjunctive query equivalence, this holds even if the target constraints are restricted to primary keys. We separate the two relaxed notions of equivalence by identifying a practically relevant class of target integrity constraints (containing functional and inclusion dependencies), for which data exchange equivalence is decidable and offers more optimization potential than logical equivalence. Finally, we consider testing conjunctive query equivalence for mappings based on Second-Order tuple generating dependencies and show undecidability of this task, under a realistic assumption that primary keys are defined over the source schema.